ON ABILITY OF A COMMUNICATION CHANNEL TO ACCOMMODATE MULTIMEDIA TRAFFIC

V. Marbukh
Wireless Communications Technologies Group
National Institute of Standards and Technology
100 Bureau Drive, Stop 8920
Gaithersburg, MD 20899-8920

Abstract

The throughput of a communication system serving bursty multimedia traffic can be characterized by the admission region: all possible sets of multimedia sources the system can accommodate given the sources statistical characteristics and Quality of Service (QoS) requirements. Since throughput depends on the system resources (bandwidth and buffering space) as well as the resource management strategies, it is natural to define the system capacity as the upper limit on the admission region over all physically feasible strategies. This paper estimates the capacity of a buffered communication channel serving multimedia traffic with QoS requirements on the maximum allowable traffic delay and loss. The buffer is assumed large enough to make the channel capacity the limiting resource. Given the set of multimedia sources within the system capacity, the paper also discusses the scheduling disciplines that provide the required QoS.

1. Introduction

Emerging multimedia networks will statistically multiplex bursty multimedia traffic sources with vastly different statistical characteristics and Quality of Service (QoS) requirements. In presence of multimedia traffic, different classes of resource (bandwidth and buffer space) management strategies can be compared on the basis of the admission region. The admission region represents all possible sets of multimedia sources the system can accommodate with the corresponding strategies, given the system resources, and QoS requirements. We consider a single buffered communication channel serving multimedia traffic with QoS requirements on maximum allowable cell loss probability and delay. Due to limited paper space we assume that the buffer is large enough to make the channel capacity (not the buffering space) the limiting resource. This situation is typical for wireless communication. We assume that J Classes of Service (CoS) are statistically multiplexed through a large buffer on a communication link with a constant bit rate C. CoS j includes K_j identical sources with the following QoS requirements:

$$(1) \qquad L_j \le L_j^{\max}, \quad d_j \le d_j^{\max}$$

where L_j and d_j are the traffic loss probability and delay respectively. Given the scheduling discipline S and QoS requirements (1), the admission region $A(S)$ represents all possible sets of multimedia users $\mathbf{K} = (K_1,..,K_J) \in A(S)$ the system can accommodate. We assume the buffer management strategy which accepts all arriving traffic into the buffer and drops a cell of class j from the buffer once the cell current delay exceeds the limit $d_j^{\max}$ to prevent wasting the channel capacity on an outdated traffic. Obviously, this buffer management strategy is optimal in terms of maximizing the admission region for a sufficiently large buffer but can be improved for small buffers.

It is natural to define the system capacity region as the upper limit of the admission region $A(S)$ over all scheduling disciplines S:

$$(2) \qquad \mathbf{A} = \bigcup_S A(S)$$

Assuming that the statistical properties of the sources are known, the paper approximates the capacity region (2) as a function of the channel bit rate C and QoS requirements (1): $\mathbf{A} = \acute{\mathbf{A}}(C; g_1^{\max},..,g_J^{\max}; d_1^{\max},..,d_J^{\max})$. The paper also discusses scheduling disciplines S which closely approach the capacity region (1). We consider the following large deviations asymptotic regime: $C \to \infty$, $L_j^{\max} \to 0$, $d_j^{\max} = O(1)$, $g_j^{\max} = -(\log L_j^{\max}) / C = O(1)$. It is known [1] that under this regime the statistical properties of the sources can be characterized by the effective bandwidths:

$$(3) \qquad a_j(s,t) = \frac{1}{st} \log E\{\exp[s x_{jk}(t)]\}$$

where $x_{jk}(t)$ is the amount of traffic generated by a source k of class j during time interval $[0,t)$. We assume that processes $x_{jk}(t)$ have stationary increments and are jointly statistically independent for $k = 1,..,K_j$, $j = 1,..,J.$ The average rate of a source of class j is $l_j = \lim_{s \to +0} a_j(s,t)$.

The paper is organized as follows. Section 2 conjectures the distributional version of the Little theorem. Section 3 approximates the capacity region in a case of uniform loss and CoS specific delay requirements: $L_j \le L^{\max}$, $d_j \le d_j^{\max}$. Section 4 approximate the capacity region in a case of CoS (1). Finally, section 5 illustrates our results in a case of $J = 2$ classes of service and Brownian traffic sources.

2. The Conservation Laws

In a case of uniform QoS requirements $L_j \le L^{\max}$, $d_j \le d^{\max}$, the corresponding capacity region $\mathbf{A^u}(C, g^{\max}, d^{\max})$ can be realized with the First In First Out ($FIFO$) scheduling. It is known [1] that the boundary of this region $\partial \mathbf{A^u}$ is closely approximated by an appropriately chosen linear hyperplane. Following [2]-[3], we may pick a point $\tilde{\mathbf{K}}$ on $\partial \mathbf{A^u}$ and obtain the tangent hyperplane to $\partial \mathbf{A^u}$ which tauches $\partial \mathbf{A^u}$ at $\tilde{\mathbf{K}}$:

$$(4) \qquad \sum_j K_j e_j(g^{\max}, d^{\max}, \tilde{\mathbf{K}}) = C$$

where a parameter $e_j(g, T, \mathbf{K})$ can be justifiably called the effective rate of a source of class j. The effective rates $e_j(g, T, \mathbf{K})$ depend on the traffic mixture $\mathbf{K}$ and can be calculated if the effective bandwidths (3) are known [1].

Consider a system that statistically multiplexes a traffic mixture $\mathbf{K} = (K_1,..,K_J)$ on a channel of capacity C through an infinite buffer. Assuming that scheduling discipline treats all sources of the same class equally, let $D_j = D_j(g, \mathbf{K})$ and $b_j = b_j(T, \mathbf{K})$ be given by $\Pr\{d_{jk} \ge D_j\} = e^{-Cg}$ and $b_j = E[b_{jk} | \; b_\Sigma \ge TC]$ where d_{jk} and b_{jk} are the delay and backlog respectively for a traffic generated by a source k of class j, and $b_\Sigma = \sum_{j,k} b_{jk}$ is the total backlog. Given a traffic mixture $\mathbf{K}$ and a pair (g, T) such that

$$(5) \qquad \Pr ob\{b_\Sigma \ge TC\} = e^{-Cg}$$

we conjecture the following inequality:

$$(6) \quad e_j(g, T, \mathbf{K}) E[d_{jk} | d_{jk} \ge D_j(g, \mathbf{K})] \ge b_j(T, \mathbf{K})$$

for any scheduling discipline. This inequality can be interpreted as a distributional version of the Little theorem [4]. If $\sum_j K_j l_j < C$, $g \to 0$, then $e_j(T, g, \mathbf{K}) \to l_j$, $D_j \to 0$, $b_j \to 0$ and, as a result, (6) reduces to the Little theorem: $l_j E[d_{jk}] = E[b_{jk}]$. Inequality (6) also holds if the traffic mixture includes only identical sources. In this particular case equality in (6) can be realized with the First In First Out ($FIFO$) scheduling discipline. There is a number of arguments which can be put forward to suggest that (6) holds (at least approximately) in a general case and the boundaries in (6) can be closely approached for all $j=1,..,J$ simultaneously with the Earliest Due Date scheduling discipline $EDD(D_1,..,D_J)$ [5].

We will use (5)-(6) to derive upper bounds on the capacity region under the large deviations asymptotic regime when $E[d_{jk} \mid d_{jk} \geq D_j] \rightarrow D_j$, and consequently, (5)-(6) take the following form:

(7) $D_j e_j(g,T,\mathbf{K}) \geq b_j$ where $T=\dfrac{1}{C}\sum_j K_j b_j$

3. The Capacity Region for Uniform Losses

Combining (4) with (7) we obtain the following system:

(8) $$\sum_{j\in\mathbf{J}} K_j D_j e_j(g,T,\mathbf{K}) \geq CT$$

(9) $$\sum_{j\in\mathbf{J}} K_j e_j(g,T,\mathbf{K}) = C$$

for any subset $\mathbf{J} \subseteq \{1,2,..,J\}$. Solving (9) for T and then substituting this T into (8) we obtain a region (a simplex exterior) $\tilde{R}_{\mathbf{J}}(\mathbf{K},C)$ in a parameter space $(g,D_1,..,D_J)$. Intersection of these regions $\tilde{R}(\mathbf{K},C)=\bigcap \tilde{R}_{\mathbf{J}}(\mathbf{K},C)$ for all subsets $\mathbf{J}\subseteq\{1,..,J\}$ upper bounds the attainable QoS region $R(\mathbf{K},C)=\{g,D_1,..,D_J\}$ in a case of uniform loss probabilities: $g_1=..=g_J=g$. Our hypotheses are that region $\hat{R}(\mathbf{K},C)$ in fact closely approximates region $R(\mathbf{K},C)$: $R(\mathbf{K},C)\cong\tilde{R}(\mathbf{K},C)$, and the QoS region can be closely approached with the $EDD(d_1^{\max},..,d_J^{\max})$ scheduling discipline.

Consider a case of uniform loss and service specific delay requirements: $L_j \leq L^{\max}$, $d_j \leq d_j^{\max}$, $j=1,..,J$. Without loss of generality assume the CoS are arranged according to the stringency of the QoS requirements: $d_1^{\max} \leq d_2^{\max} \leq .. \leq d_J^{\max}$. All traffic mixtures $\mathbf{K}$ for which the attainable QoS region (8)-(9) contains the QoS parameters $L_i = L^{\max}$, $d_i = d_i^{\max}$, $i\in\{1,..,j\}$ form the following region: $U_j=\{K_i \mid \sum_{i\leq j} K_i e_i(g^{\max},T_j,K_1,..,K_j)\leq C\}$ where $T_j=T_j(K_1,..,K_j)$ is the solution to the following fixed point equation: $T_j=\dfrac{1}{C}\sum_{i\leq j}K_i d_i^{\max} e_i(g^{\max},T_j,K_1,..,K_j)$. Regions U_j, and, consequently, their intersection $\mathbf{U}=\bigcap U_j$, $j=1,..,J$ upper bound the capacity region (2) in a case of uniform losses

(10) $$\mathbf{A}(C,g^{\max},d_1^{\max},..,d_J^{\max})\subseteq\mathbf{U}$$

Upper bound (10) is obtained by considering the attainable QoS regions (8)-(9) for J subsets $\{1,..,j\}$, $j=1,..,J$ of the set of services $\{1,..,J\}$. It can be shown that applying conservation laws to *all possible* subsets of the set of services $\{1,..,J\}$ *would not produce a tighter upper bound* on the capacity region than (10). We expect that the intersection $\mathbf{U}$, in fact, closely approximates the capacity region (2) for uniform losses:

(11) $$\mathbf{A}(C,g^{\max},d_1^{\max},..,d_J^{\max})\cong\mathbf{U}$$

and that the capacity region (11) can be closely approached with the $EDD(d_1^{\max},..,d_J^{\max})$ scheduling.

4. The Capacity Region in a General Case

First, consider a "homogeneous "case when the traffic mixture consists only of sources of the same class j: $K_i=0,\ i\neq j$. Given the QoS requirements (1), the corresponding capacity region $\mathrm{Á}_{\mathrm{j}}^{\mathrm{h}}(C,g_j^{\max},d_j^{\max})$ is $K_j \leq K_j^{\max}=C/e_j^h(C,g_j^{\max},d_j^{\max})$ where e_j^h is the corresponding "homogeneous" effective rate. Consider all pairs (g,d) for which: $\mathrm{Á}_{\mathrm{j}}^{\mathrm{h}}(C,g,d)=\mathrm{Á}_{\mathrm{j}}^{\mathrm{h}}(C,g_j^{\max},d_j^{\max})$, or, equivalently,

(12) $e_j^h(C,g,d) = e_j^h(C,g_j^{\max},d_j^{\max})$

Equation (12) describes the trade-off between the QoS requirements on loss (g) and delay (d) in a homogeneous case: $K_i = 0,\ i \neq j$. Assuming that this trade-off holds in a heterogeneous case, when a mixture of different services is multiplexed on the channel, one may try to approximate the heterogeneous capacity region as follows:

$$(13)\qquad \mathbf{A}(C;g_1^{\max},..,g_J^{\max};d_1^{\max},..,d_J^{\max}) \cong$$

$$\cong \mathbf{A}(C;g;d_1,..,d_J)$$

where region $\mathbf{A}(C;g;d_1,..,d_J)$ is approximated according to (11), and $d_j = d_j(g;g_j^{\max},d_j^{\max})$ are determined by (12) for some g. After implementing this procedure for general traffic sources the resulting approximate capacity region will depend on the arbitrary parameter g since the trade-off between the loss and delay depends on the traffic mixture. However, for Brownian traffic sources [1] the effective rate is $e_j(g,d) = l_j + \frac{s_j^2}{2}\frac{g}{d}$, equation (12) takes the following form: $d/g = d_j^{\max}/g_j^{\max}$ and, consequently, the right-hand side of approximation (12) is independent of the parameter g. Arrange services as follows: $w_1 \leq .. \leq w_J$, where $w_j = d_j^{\max}/g_j^{\max}$. Using (11)-(13) we can approximate the capacity region in a case of QoS requirements (1) and Brownian traffic sources as follows:

$$(14)\qquad \mathbf{A}(C;g_1^{\max},..,g_J^{\max};d_1^{\max},..,d_J^{\max}) \cong \bigcap_{j=1}^{J} U_j$$

where regions U_j are as follows: $\{K_i \mid \sum_i K_i l_i + \frac{1}{2c_j}\sum_i K_i s_i^2 \leq C,\ i = 1,..,j\}$ and c_j are easily obtained from the following quadratic equations:

$$c_j = \frac{1}{C}\sum_{i \leq j} K_i w_i\, l_i + \frac{1}{2Cc_j}\sum_{i \leq j} K_i w_i\, s_i^2$$

We expect that for Brownian sources approximation (14) is in fact an identity and for general traffic sources (14) closely approximates the capacity region. We also expect that for Brownian (arbitrary) sources the capacity region can be realized (closely approached) with the scheduling discipline $S(g_1^{\max},..,g_J^{\max};w_1^{\max},..,w_J^{\max})$. This discipline can be interpreted as a generalization of the EDD with service j specific time scale: $t \rightarrow t_j = t/g_j^{\max}$ and assuming that service j specific deadline with respect to the time scale t_j is $w_j^{\max}$.

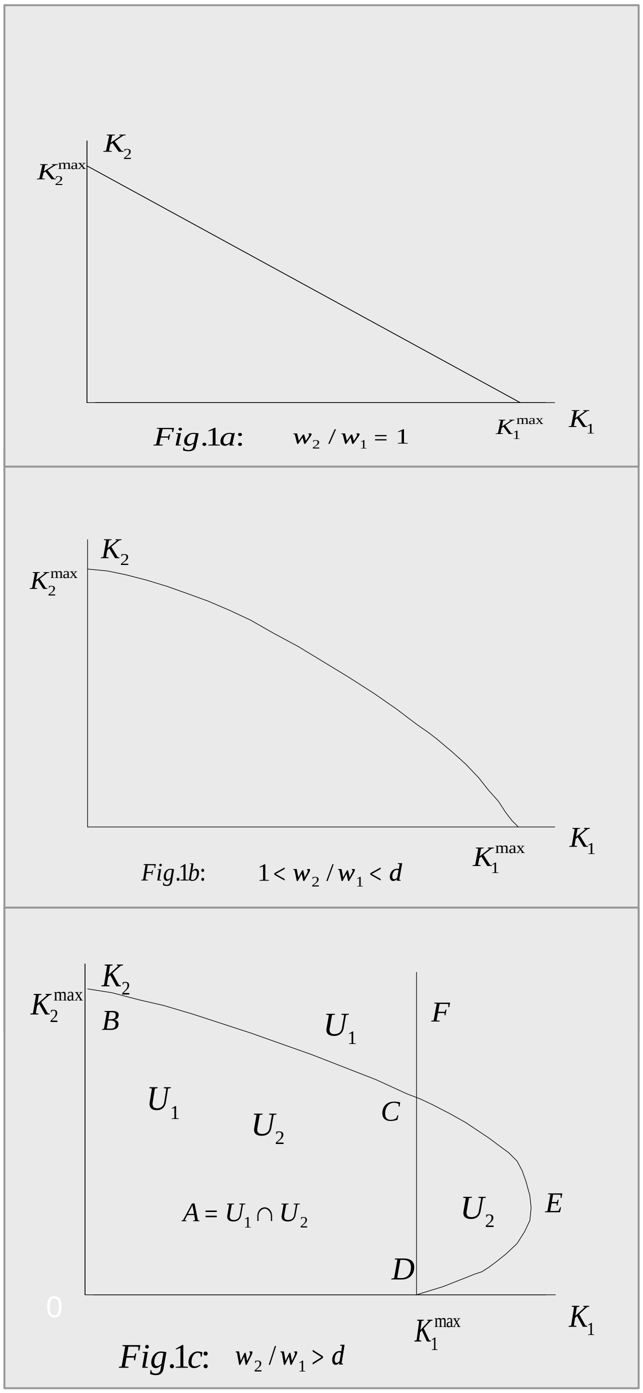
It can be shown that the region U_j can be closely approximated by an appropriately chosen simplex: $\{K_i \mid \sum_i K_i e_{ji}^* \leq C,\ K_i \geq 0,\ i = 1,..,j\}$, and, consequently, the capacity region (13) can be approximated by the J-dimensional polyhedron:

$\{K_j \mid \sum_{i \leq j} K_i e_{ji}^* \leq C,\ K_j \geq 0,\ \ j = 1,..,J\}$. This approximation is consistent with theoretical and simulation results for specific scheduling disciplines [6]-[7].

5. Example: Two Classes of Service, Brownian Traffic Sources

As an example consider a case of $J = 2$ classes of service, and Brownian traffic sources. We use the same notations as in the previous section and assume that $w_1 \leq w_2$. Fig. 1 shows regions U_1, U_2 and their intersection $\mathbf{A} = U_1 \bigcap U_2$ in the following three cases: (a) $w_2/w_1 = 1$, (b) $1 < w_2/w_1 < d$, and (c) $w_2/w_1 > d$ where

$$(15)\qquad d = 2(1 + 2\frac{l_1 w_1}{s_1^2})$$



$Fig.1a$: $w_2 / w_1 = 1$

$Fig.1b$: $1 < w_2 / w_1 < d$

$Fig.1c$: $w_2 / w_1 > d$

Figure 1: Schematic view of the capacity region in a case of two classes of service.

If $w_2 / w_1 = 1$, then the capacity region $\mathbf{A} = U_2 \subset U_1$ has linear boundary since this case in effect is a case of uniform QoS requirements. If $1 < w_2 / w_1 < d$, then still $\mathbf{A} = U_2 \subset U_1$, but the capacity region has non-linear boundary. If $w_2 / w_1 > d$ then $U_2 \not\subset U_1$, and the boundary (DCB) of the capacity region $\mathbf{A}$ contains a linear segment (DC). The trapezoidal shape of the capacity region in a case $w_2 / w_1 > d$ indicates that even when the system is serving the maximum possible number of high priority sources of class 1, the channel still has enough residual capacity to accommodate certain number of sources of class 2 as low priority. Examination of (15) shows that this is possible if the high priority sources are burty and the difference in the QoS requirements for different services is sufficiently large to allow for a squeezing of the low priority traffic between the bursts of the high priority traffic.

References

1. F.P. Kelly, "Notes on effective bandwidth," *Stochastic Networks*, F.P. Kelly, S. Zachary, I. Ziedins, Eds., Oxford: Clarendon Press, 1996, pp. 141-68.

2. J.Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks*, Boston, MA: Kluwer, 1990.

3. F.P. Kelly, "Effective bandwidth at multi-type queues," *Queueing Syst.*, vol. 9, pp. 5-15, 1991.

4. W. Whitt, "A review of $L = lW$," *Queueing Systems*, vol. 9, 1991, pp. 235-68.

5. S.S Panwar, D. Towsley, and J.K. Wolf, "Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service," *J. Ass. Comput. Mach.*, pp. 832-844, 1988.

6. A.W. Berger and W. Whitt, "Extending the effective bandwidth concept to networks with priority classes," *IEEE Comm. Magazine*, vol. 36, no.8, Aug. 1998, pp.78-83.

7. A.I. Elwalid and D. Mitra, "Analysis, approximations and admission control of a multi-service multiplexing system with priorities," *Proc. IEEE INFOCOM'95*, pp. 463-72, 1995